Public Transp
DOI 10.1007/s12469-016-0138-7

ORIGINAL PAPER

Train unit scheduling guided by historic capacity provisions and passenger count surveys

Zhiyuan Lin[1] · **Eva Barrena**[1] · **Raymond S. K. Kwan**[1]

Accepted: 25 August 2016

Abstract Train unit scheduling concerns the assignment of train unit vehicles to cover all the journeys in a fixed timetable. Coupling and decoupling activities are allowed in order to achieve optimal utilization while satisfying passenger demands. While the scheduling methods usually assume unique and well-defined train capacity requirements, in practice most UK train operators consider different levels of capacity provisions. Those capacity provisions are normally influenced by information such as passenger count surveys, historic provisions and absolute minimums required by the authorities. In this paper, we study the problem of train unit scheduling with bi-level capacity requirements and propose a new integer multicommodity flow model based on previous research. Computational experiments on real-world data show the effectiveness of our proposed methodology.

Keywords Train unit scheduling · Required train capacities · Multicommodity network flow

1 Introduction

A train unit is a self-propelled, relatively small fixed set of rolling stock carriages (or *cars*) that can move in either track direction on their own, in contrast to a traditional configurable combination of locomotive(s) and cars with the locomotive as the only power source. This is the most commonly used passenger rolling stock in

✉ Zhiyuan Lin
z.lin@leeds.ac.uk

Eva Barrena
e.barrena@leeds.ac.uk

Raymond S. K. Kwan
r.s.kwan@leeds.ac.uk

[1] School of Computing, University of Leeds, Leeds LS2 9JT, UK

Published online: 08 September 2016

the UK and many other European countries. A timetable is a set of train services/trips, conventionally called *trains*, during one working day, each of which has attributes mainly consisting of departure and arrival stations and times, seat demand, coupling and decoupling constraints and allowed types of train units. Given a fixed timetable on one operational day and a fleet of train units of multiple types, the train unit scheduling problem (TUSP) (Lin and Kwan 2013, 2014) aims at deriving an optimized plan such that all the trains are covered with the required seat capacity provisions. From the perspective of a train unit, the problem assigns a sequence of trains to it as its daily workload. A notable feature of the TUSP is the activity of unit coupling/decoupling in response to different passenger demands. Generally, a train with a high demand may require coupled units. In addition, coupling can also be used as a way of redistributing unit resources across the rail network regardless of the demand en route. Similar or relevant problems with respect to the TUSP include train unit circulation (Schrijver 1993; Alfieri et al. 2006; Fioole et al. 2006; Peeters and Kroon 2008) and train unit assignment (Cacchiani et al. 2010, 2012, 2013b).

Common objectives in the TUSP include minimizing the number of units used, carriage-mileage and number of empty-running trains. Constraints regarding the number of coupled units are also required. While coupled units may be needed to provide sufficient seat capacity, the number of coupled cars must not exceed a limit that can depend on routes and/or unit types. Other constraints include aspects such as unit coupling compatibility relations among traction types, locations banned for coupling/decoupling, and unit blockage prevention.

Most of the relevant research in passenger rolling stock scheduling in the literature considers a single level of capacity provision requirements. Those requirements may not only depend on a single aspect such as passenger demands, but are also influenced by other factors such as historic capacity provisions and robustness. Solely relying on passenger count surveys may not be appropriate since, for example, fluctuations on passenger demand may lead to low robustness in the resulting schedules. On the other hand, it may not be correct to infer capacity requirements solely from historic schedule because excessive or insufficient provisions might have resulted from scheduling logistics in the past that are no longer relevant. When an "optimized" schedule has some train units with very little work assigned, it may be appropriate to utilize such train units to provide extra capacity on some targeted trains. It is therefore insufficient to include a single level of capacity requirements in the scheduling model.

In this paper, we propose to incorporate two levels of capacity requirements, namely a target (lower) level that has to be satisfied strictly and a desirable (higher) level that is to be achieved as much as possible. In doing so, we guarantee the capacity provision at the target level and minimize deviation from the desirable level, while using the minimum fleet size and mileage. An integer multicommodity flow model for train unit scheduling based on previous work in Lin and Kwan (2013, (2014) is proposed such that the bi-level capacity requirements will be considered. The model strictly satisfies the target capacity requirement as integer linear programming (ILP) constraints while it tries to achieve the desirable capacity requirement through the objective function.

The remainder of this paper is organized as follows. In Sect. 2 we survey the relevant research in train unit resource planning in the literature. In Sect. 3 we describe the specific problem under consideration, as well as the reason why there is a need for a bi-level capacity model. Section 4 describes the model formulation and solution algorithm. Finally, in Sect. 5, we present some computational experiments based on real datasets from First ScotRail.

2 Literature review

The TUSP, particularly for the problem scenarios in the UK, has been studied in Lin and Kwan (2013, (2014). A branch-and-price ILP solver has been designed to solve the problem exactly for up to 500 train instances. Many real-world objectives and constraints that were ignored in previous studies are considered, e.g. unit type coupling compatibility, locations banned for coupling/decoupling, time consumption due to coupling/decoupling, and elimination of unnecessary coupling/decoupling. Moreover, in Lin and Kwan (2014), a two-phase approach is proposed where the first phase as an integer fixed-charge multicommodity flow model assigns and sequences train trips to the fleet temporarily ignoring some station infrastructure details, and the second phase performs post-processing to satisfy any remaining detailed requirements at each station. Although in Lin and Kwan (2014) the post-processing is modeled as a multidimensional matching problem, currently TRACS-RS (Tracsis PLC 2015), a software package that aims at facilitating human schedulers' manual process by visualizing and resolving blockage and shunting plans at individual stations, is used to perform the second phase interactively.

The train unit assignment problem (TUAP) (Cacchiani et al. 2010) shares very similar definitions and settings with the TUSP, in particular no trains/trips are pre-sequenced in advance. The TUSP considers additional aspects such as location banned for coupling/decoupling, unit type compatibility and combination specific coupling upper bounds. Cacchiani et al. (2010) present an integer multicommodity flow model for the TUAP which is based on a directed acyclic graph similar to the one to be used in Sect. 4 and a path formulation ILP based on the graph is used. Noting that tested instances have a feature that no more than two units can be coupled, relevant knapsack constraints are strengthened by describing their dominants explicitly. An LP-based diving heuristic is designed for finding the integer solutions. This heuristic can solve problem instances of up to 600 trains. In addition, Cacchiani et al. (2013a) give proofs on explicit convex hull descriptions for the knapsack polytopes to strengthen the weak LP relaxation, which have been implemented into the models in Cacchiani et al. (2010). A similar approach in describing convex hulls is used in the model in this work as well.

The train unit circulation problem (Schrijver 1993; Maróti 2006) simultaneously plans over multiple days. There have been extensive studies in this area and they are applied to real-world instances mainly at NSR, a Dutch passenger train operator. Each train may be identified with one predecessor and one successor in advance, where cyclic timetables are practised. Schrijver (1993) proposes early work on this problem with two types of train units. Alfieri et al. (2006) further extended the

above work with two models where the first one uses a normal multicommodity flow framework without considering unit permutations while the second one makes use of transition graphs to handle unit permutations. Peeters and Kroon (2008) further developed a branch-and-price solver for similar problems as in Alfieri et al. (2006) to give exact solutions for real-world instances. Fioole et al. (2006) consider a special scenario of combining and splitting trains.

Other relevant research on train unit planning/scheduling include the following. Cadarso and Marín (2011) consider passenger rolling stock scheduling with stochastic passenger demands; Fuchsberger and Lüthi (2007) solve the train scheduling problem in a main station area using a resource-constrained space-time integer multi-commodity flow model; Kroon et al. (2008) deal with train units shunting in a large station with complex track layouts; Jiang et al. (2014) discuss scheduling additional train unit services on Shanghai rail transit line 16.

All the above-mentioned researches consider a single level of capacity requirements. In fact, to the best of our knowledge, none of the existing works in the literature deal with two-level capacity requirements, which is the main focus of this paper.

3 Problem description

3.1 Train capacity requirement information

Each train in a timetable should be covered by a unit or coupled units whose total capacity satisfies the passenger demand expected for the train, which is usually measured in number of seats. For the TUSP, train capacity requirements are very important, due to its significant impact on objectives such as fleet size and unit resource distribution pattern over the rail network. On the other hand, in the UK rail industry capacity requirement information is usually patchy and lacking documentation, making it difficult to be determined precisely.

Let N be the set of train trips. At First ScotRail, the major train operator in Scotland, passenger capacity requirement information for each train service $j \in N$ in a new timetable can be mainly inferred from three sources, which will be referred to as "raw data":

(i) Mandatory minimum capacity ρ_j^M: The mandatory minimum capacity is required by the authorities or franchise agreements. In principle, it must be satisfied as a bare minimum level of capacity provision.

(ii) Historic capacity provisions ρ_j^H: Capacity provisions given by operator's schedules operated in the past are available for reference. Since a large proportion of trains will remain unchanged in a half-yearly new timetable release, their historic capacity provisions would still be largely relevant.

(iii) Passenger count surveys (PAX) ρ_j^P: Every year, a subset of trains will have their actual on-board passenger numbers counted, which is referred to as "PAX".

For each train, its PAX can be compared with historic capacity. We say that a train is *over-provided* (OP) if its historic capacity exceeds its PAX in the sense that this PAX could be satisfied with a train unit combination of smaller capacity than the historic one. For example, if there is only one type of train unit, OP would mean that its historic capacity exceeds its PAX in terms of unit numbers (i.e. x unit(s) would be sufficient for its PAX but the historic schedule uses at least $x + 1$ units). OP trains may be caused by, e.g., a lack of available place for decoupling. Another reason is that excessive capacity provision may be a by-product of relocating unit resources to satisfy trains later elsewhere. Finally OP trains may be merely a result of an under-optimized unit schedule. On the other hand, a train is *under-provided* (UP) if its historic provision fails in satisfying its PAX. Such under-provision is more likely to occur during peak hours when demands are much higher in many locations across the network while the fleet size and the maximum numbers of coupled units are both limited.

The raw data from the above three sources may not be complete or accurately reflecting the "ideal" capacity provision level a rail network requires. The mandatory minimum level is generally too low for practical schedules and thus can only be used as a basic lower bound. The other two sources are discussed in the following.

3.1.1 Historic capacity provisions

Historic capacity provisions often contain useful information on the basic pattern of unit resource distribution over a network, such as the unit capacity provision per train service and how different services are connected. as well as implicit agreements or understanding with transport authorities. Nevertheless, simply applying them to a new timetable will not be reliable and sufficient, even assuming most trains remain unchanged.

In historic capacity records, many of the strengthened capacities achieved by coupling are in fact used to redistribute unit resources over the network rather than satisfying real demands on the trains concerned. Thus they may be unnecessary in an updated timetable and train unit schedule. Moreover, even excluding the unit redistribution factor, historic records still may not be flawless in reflecting true capacity requirements. The manual process in train unit scheduling is basically modifying previous schedules, subject to changed parts in a new timetable in a station-by-station manner, leaving the backbone of a new schedule heavily similar to previous ones. Therefore, if there were unreasonable patterns in previous schedules, they are likely to be passed down to a new schedule year after year without being challenged or reconsidered.

3.1.2 PAX surveys

Although PAX surveys reflect the real passenger numbers, directly using them as capacity requirements may not be realistic, not only because merely a subset of trains is surveyed, but also due to issues like robustness and limited fleet size that cannot satisfy all UP trains.

For some instances, the overall PAX level can be much lower compared with historic capacities yielding many OP trains. Simply reducing the capacity provision of all OP trains from historic records to PAX may affect the robustness of services. Moreover, resulting schedules may include underused units, e.g. units only serving one or two trains as their daily workload because of the minimization of carriage-mileage. By appropriately keeping the capacity requirements for some OP trains at their (higher) historic schedule level, the underused train unit resources may be assigned to cover more trains, which makes the overall schedule more balanced. Therefore it is reasonable to adjust the capacity requirements to have some of the OP trains to set their capacity requirements as historic and the others as PAX. However which subset of trains should be so adjusted is unclear.

On the other hand, for some instances the PAX levels for peak hour trains are too high such that the appearance of many UP trains is inevitable given a limited fleet size. Nevertheless, a subset *SP* of the UP trains can be identified to increase their capacity requirements from historic to PAX without violating the fleet size bound. However, it is also unclear how to decide which subset of the UP trains to strengthen.

Finally, it is possible that both OP and UP trains are present in manual schedules, making the problem more complicated as the two can be conflicting with each other with limited unit resources.

4 Model and formulation

This paper proposes a novel TUSP integer multicommodity flow model that can achieve appropriate capacity provisions considering two levels of capacity requirements derived from raw capacity information such as capacity provisions in past operated schedules and PAX.

The proposed bi-level capacity requirement model is derived from the models in Lin and Kwan (2013, (2014). It is based on a directed acyclic graph (DAG) $\mathcal{G} = (\mathcal{N}, \mathcal{A})$, where $\mathcal{A} = A_0 \cup A$ denotes the arc set and $\mathcal{N} = \{s, t\} \cup N$ denotes the node set, with s and t being the source and sink respectively and N being the set of all timetabled trains. $A_0 = \{(s,j) : j \in N\} \cup \{(j,t) : j \in N\}$ is the set of sign-on/-off arcs, where a sign-on arc links the source to a train node and a sign-off arc links a train node to the sink. Every train node has a sign-on arc and a sign-off arc assigned. A denotes the set of connection-arcs where a connection-arc $a = (i,j) \in A$ links two train nodes i and j if it is feasible for i and j to be served consecutively by the same train unit. P is used to denote the set of all s-t paths in $\mathcal{G}$ such that each $p \in P$ represents a sequence of trains as a workload plan for a unit. Moreover, P_j is used to denote the set of paths passing through node j.

As for the fleet, let K be the set of unit types, corresponding to the commodities in a multicommodity flow model. Type-graphs $\mathcal{G}^k = (\mathcal{N}^k, \mathcal{A}^k)$ as sub-graphs of $\mathcal{G}$ are constructed with respect to each type $k \in K$ generally based on the principle that a type-graph $\mathcal{G}^k$ will only contain train nodes $\mathcal{N}^k$ (apart from s, t as mandatory) that are compatible with units of type k (and arcs $\mathcal{A}^k$ to be constructed accordingly). The

components of $\mathcal{G}^k$ will also be denoted in a similar way, e.g. P^k represents the set of paths in $\mathcal{G}^k$.

The model can be formulated as an arc-based or path-based problem. In order to use column generation procedure to solve this problem, the path-based formulation is required. Moreover, according to Cacchiani et al. (2010), much shorter time is needed to solve the problem using path variables than arc variables. Based on this, we only have tested the path formulation in our experiments. We therefore present the following two kinds of decision variables:

- $x_p \in \mathbb{Z}_+, \forall p \in P^k, \forall k \in K$ represent the number of type-k units used for a path p in $\mathcal{G}^k$.
- $y_j \in \mathbb{R}_+, \forall j \in N$ represent the capacity provision for train j.

The first level of capacity requirements is a target capacity $r_j, \forall j \in N$, that must be satisfied. The second level of capacity requirements is a desirable capacity $r'_j, \forall j \in N$, that will be satisfied as much as possible but which is not mandatory. How to convert raw data to the two levels of capacity requirements will be problem-specific. A basic rule would be to ensure that $r_j \leq r'_j$. For example, $r_j = \min(\rho_j^H, \rho_j^P)$ and $r'_j = \max(\rho_j^H, \rho_j^P)$. In this paper, all train capacities are measured in number of seats.

Figure 1 illustrates how different capacities are processed within the model. The raw data such as historic capacity provision and PAX will be converted into two levels of capacity requirements—a lower target capacity and a higher desirable capacity.

To satisfy target capacity requirements $r_j, \forall j \in N$, and the requirement on the maximum number of units when coupled (see Sect. 1), an enumeration on all possible unit combinations is made for each train service (Lin and Kwan 2014). Let K_j be the set of permitted types for train j, and let $w^j = (w_1^j, w_2^j, \cdots, w_{|K_j|}^j)^T \in \mathbb{Z}_+^{K_j}$ be a unit combination at j where w_k^j stands for the number of units of type k used for j. A unit combination set $W_j, \forall j \in N$ is defined as:

$$W_j := \left\{ w^j \in \mathbb{Z}_+^{K_j} \middle| w^j \text{ is a feasible unit combination for train } j \right\}, \tag{1}$$

where the feasibility of unit combinations is given by:

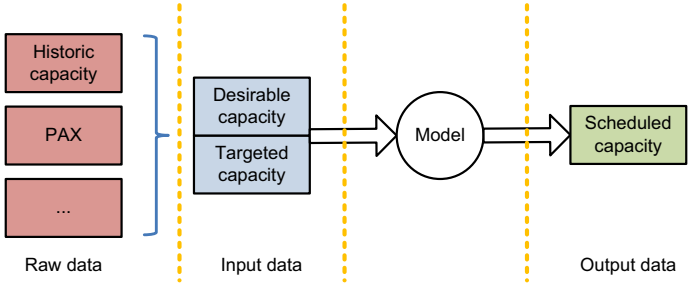


Fig. 1 Flow-chart of capacity requirements treatment in our model

(i) $\sum_{k\in K_j}\sum_{p\in P_j^k} q_k x_p \geq r_j$, i.e. the target capacity requirement r_j is strictly satisfied for each train $j \in N$, where q_k is the unit capacity of type k in number of seats.
(ii) A unit combination assigned to j is within its coupling upper bound.
(iii) The used unit types for j are compatible.

Then the corresponding train convex hulls are computed based on unit combination sets as

$$\mathrm{conv}(W_j) = \left\{ w^j \in \mathbb{R}_+^{K_j} \middle| H^j w^j \leq h^j \right\}, \forall j \in N, \tag{2}$$

which is described by non-zero facets $f \in F_j$ such that $H^j \in \mathbb{R}^{F_j \times K_j}$ and $h^j \in \mathbb{R}^{F_j}$. Via variable conversion $w_k^j = \sum_{p\in P_j^k} x_p$, the passenger demand and coupling upper bound requirements at train j can be satisfied by the following train convex hull constraints

$$\sum_{k\in K_j}\sum_{p\in P_j^k} H_{f,k}^j x_p \leq h_f^j, \quad \forall f \in F_j, \forall j \in N. \tag{3}$$

Incorporating the above train convex hull constraints, we propose the ILP formulation (P) on the integer multicommodity flow model for the TUSP with two levels of capacity requirements as

$$(P) \quad \min \quad C_1 \sum_{k\in K}\sum_{p\in P^k} c_p x_p + C_2 \sum_{j\in N} \left| y_j - r_j' \right| \tag{4}$$

$$\text{s.t.}(3) \text{ and } \quad \sum_{p\in P^k} x_p \leq b^k, \qquad \forall k \in K; \tag{5}$$

$$\sum_{k\in K_j}\sum_{p\in P_j^k} q_k x_p = y_j, \qquad \forall j \in N; \tag{6}$$

$$x_p \in \mathbb{Z}_+, \quad \forall p \in P^k, \forall k \in K. \tag{7}$$

The first term in the objective function (4) is the sum of all the used paths' costs where c_p is the weighted cost for path p with sub-weights on different components. An overall weight C_1 is set for it. Typically, c_p includes sub-terms with respect to fleet size, carriage-mileage, empty-running movements, and preferences. Specifically, in this work we set $c_p = C^{FS} c_p^{FS} + C^{CM} \sum_{a\in A_p} c_a^{CM} + C^{ER} \sum_{a\in E_p} c_a^{ER}$, where c_p^{FS} is the fleet size cost for using one unit; C^{FS} is the sub-weight on fleet size; c_a^{CM} is the carriage-mileage cost implied by arc a formulated with preferences regarding type-route, maintenance gap and so on; A_p is the set of arcs in path p; C^{CM} is the sub-weight on carriage-mileage; c_a^{ER} is the cost of an empty-running movement when arc a implies such a movement; E_p is the set of empty-running arcs in path p and C^{ER} is the empty-running sub-weight. In our experiments, we use a simplified

setting of $c_a^{CM} = 1$ for all arcs' carriage-mileage costs. Therefore, regarding carriage-mileage, we will simply report the number of used arcs in the experiment section. The second term in (4) is the sum of deviations between the desirable capacity and the solution's real provision with a weight C_2. In what follows, we call the first term the "path cost term" and the second term the "OP deviation term".

Besides Constraints (3) as aforementioned, Constraints (5) ensure that the deployed unit number per type k will not exceed its fleet size limit b^k. Constraints (6) define the solver's capacity provision for each train. Finally, Constraints (7) give the variable domains.

To overcome the non-linearity caused by the absolute value expression in the objective function and to convert (P) into an ILP, a conventional remedy is used. We create a pair of variables $y_j^+, y_j^-, \forall j \in N$ and take the replacement $|y_j - r'_j| = y_j^+ + y_j^-$ and $y_j - r'_j = y_j^+ - y_j^-$, $\forall j \in N$ in the original model. Therefore, in the actual formulation, the OP deviation term in the objective function (4) becomes $C_2 \sum_{j \in N}(y_j^+ + y_j^-)$ and Constraints (6) become $\sum_{k \in K_j} \sum_{p \in P_j^k} q_k x_p = y_j^+ - y_j^- + r'_j, \forall j \in N$.

Compared with the models in Lin and Kwan (2013, (2014), (P) has removed the "fixed-charge" components (i.e. the binary variables based on arcs indicating whether an arc is used or not), making it a standard integer multicommodity flow problem. This significantly improves the efficiency of the solution process. Furthermore, the remaining tasks to be achieved by the fixed-charge components in eliminating excessive coupling/decoupling and ensuring connection time allowance involving coupling/decoupling can be handled by post-processing as mentioned in Sect. 2 after solving the main ILP model. TRACS-RS (Tracsis PLC 2015), a visualization tool for rolling stock scheduling, is currently used for post-processing. At each station, the connection relations between arrival and departure trains can be manually adjusted as indicated by the graphical visualization to satisfy the aforementioned requirements and to avoid any violation on operational rules. As the focus of this paper is on the bi-level capacity requirements, we choose to not include the fixed-charge terms in (P). Similar strategies in achieving the bi-level requirements can be applied to the full version with fixed-charge components by analogy.

To solve (P) exactly, a similar branch-and-price method as in Lin and Kwan (2013, (2014) is used. The paths are dynamically generated by shortest path subproblems per unit type. Since the "fixed-charge" terms are no longer used in (P), the model only considers the path variable x_p. As for BB tree node traversing, an adaptive strategy combining best-first and depth-first was used for all runs. Two customized branching methods named banned location branching and train-family branching are embedded into the relevant branch-and-bound (BB) tree. Banned location branching will identify LP-relaxation solutions at BB tree nodes with coupling/decoupling operations at locations banned for these activities and form branches to gradually remove them. Train-family branching will identify LP-relaxation solutions at BB tree nodes with coupling-incompatible unit types covering the same train and form branches to allow only compatible types at each

child node. Appropriate post-processing on a station-by-station basis is used to eliminate excessive coupling/decoupling and remove unit blockage, yielding a finalized operable solution for train operating companies.

5 Computational experiments

Our work is based on real-world data provided by First ScotRail. Two groups of experiments were conducted. The first group is on a 2011 timetable instance in which all PAX is satisfied with a large proportion of over-provided trains. The focus for this instance is to find a relation between the capacity provision level and fleet size such that the operator is able to find a trade-off point based on its own needs. The second group of experiments was conducted on a 2013 timetable instance in which not all PAX is satisfied, that is, there exist under-provided trains. The focus here is to minimize the number of under-provided trains found in the manual schedule while maintaining the same fleet size.

5.1 Rail network and historic schedule

We have performed experiments based on the datasets of the Central Scotland railway network (see Fig. 2). First ScotRail normally divides the Central Scotland network management into two areas: North and South. For the North area, we will consider the December 2011 operated schedule and for the South area the December 2013 operated schedule. We focus on the capacity provision and need a fast execution solver for experiments. For this reason and simplicity, we have considered just one type of train unit in each of the following experiments. From

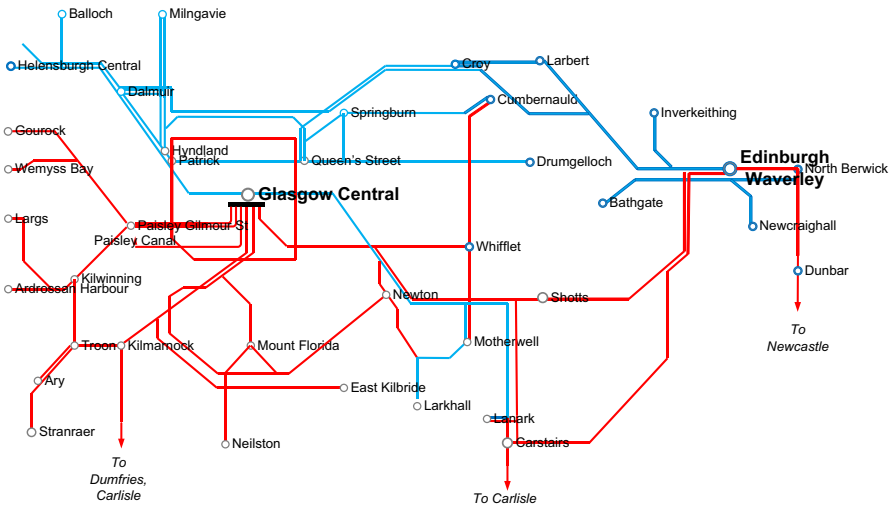


Fig. 2 Central Scotland railway network (with the North area routes in *blue* and the South area routes in *red*) (Color figure online)

now on, we will call *OP* and *UP* the set of over-provided and under-provided train services, respectively, by comparing the historic schedule and the PAX. Table 1 gives a summary of the problem instances extracted for each train unit type, as well as the *OP* and *UP* trains in the schedules operated by First ScotRail. So, for example, in the operated schedule provided, all the demand of each of the 156 trains in the North area served by train units of Class 334 was satisfied by means of 33 train units, which results in 64 over-provided train services. However, there are 12 under-provided trains in the south area served by train units of Class 314, that is, not all demand was satisfied for those trains.

The experiments were conducted using a 64 bit Xpress-MP 7.7 package on a workstation with Intel Core i7-4790 CPU.

5.2 Experiments on north area c334 instance

Observe that the terms in the objective function (4) are competing. Minimizing the OP deviation term implies augmenting the fleet size and/or the current carriage-mileage (simplified to the number of used arcs in the experiments), which are part of the path cost term. The weights of the terms in the objective function will then have a great impact on the resulting schedules and its calibration becomes an important issue.

First, in Sect. 5.2.1 we show the results by varying the weights of the objective function terms dealing them with the bi-objective problem by means of the weighted sum method. We observe that the same fleet size may over-provide a different number of trains. Second, in order to obtain the maximum number of *OP* trains that can be achieved within a certain fleet size, in Sect. 5.2.2, we make the use of the ε-constraint method. Specifically, we fix parametrically an upper bound for the fleet size and aim to minimize the deviation w.r.t. *OP* trains in the existing schedule.

Table 1 Summary of problem instances and *OP* and *UP* trains in the December 2011 and 2013 operated schedule

	North area	South area
Train unit type	Class 334	Class 314
Number of origin/destination stations (among which coupling/decoupling is banned)	11 (6)	7 (2)
Operational period	One working day	One working day
Fleet size	33	14
Number of train services	156	278
Number of OP trains	64	9
Number of UP trains	0	12

5.2.1 Calibrating objective function weights

(i) Single-type case

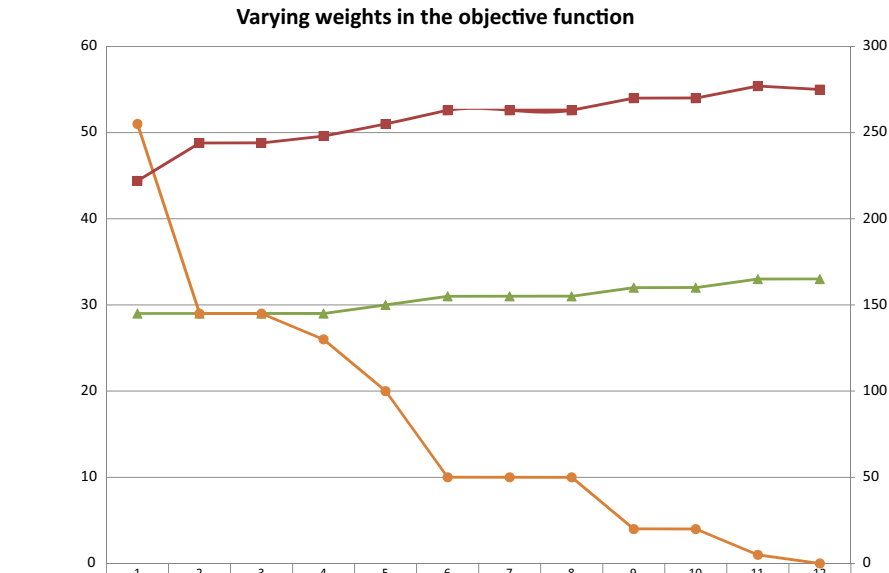
In this set of experiments, we vary the weights C_1 and C_2 in (4), where $C_1 + C_2 = 1$, and observe the impact of them in the resulting train unit schedules. For that purpose, we gradually increase C_2 and therefore C_1 will decrease accordingly, thus yielding a higher number of *OP* trains. Results are presented in Table 2 and graphically depicted in Fig. 3. In Table 2, ECS# gives the number of empty-running trains produced by the model.

It can be observed that, as expected, the fleet size tends to increase in order to reduce the *OP* deviation (measured in numbers of trains in the table). On the other hand, the same fleet size may yield different number of *OP*, e.g. rows 1–4 in Table 2, the same fleet size of 29 train units leads to different *OP* deviation in the interval from 26 to 51. In the fourth column it can be seen that the number of used arcs also increases when one aims to over-provide more trains within the same fleet size. This is affected by the fact that the same fleet size incur higher mileage in order to over-provide more trains.

Comparisons are made between the results of our model and those of the historic schedule in which 33 train units are required to satisfy the demand of all train services with 64 over-provided trains against the PAX (which is equivalent to the case when $C_2 = 1$ in our experiments). The most important issue for the train operator is to minimize the fleet size while meeting all passenger demands and having as little deviation as possible from historic capacity provisions. According to these interests and the model results, the train operator is likely to select the option with the minimum fleet size achieving the maximum possible number of *OP* trains, that is, 29 train units and 38 *OP* trains corresponding to 26 *OP* deviation (fourth row in Table 2). The fleet size in our best result is considerably reduced by 4 units w.r.t.

Table 2 Varying weights in the objective function

C_2	LP gap	Fleet size	Arcs#	OP deviation	ECS#	Time (s)	BBNode#
0	0.03	29	222	51	1	62	98
0.02	0.3	29	244	29	1	49	54
0.05	0.8	29	244	29	1	37	69
0.1	0.63	29	248	26	1	1977	1562
0.13	1.55	30	255	20	1	51	71
0.14	0.56	31	263	10	1	392	613
0.15	0.39	31	263	10	1	124	167
0.16	0.22	31	263	10	1	60	63
0.17	0	32	270	4	1	60	38
0.18	0	32	270	4	1	53	31
0.5	1.48	33	277	1	2	38	28
1	0	33	276	0	2	37	27



	1	2	3	4	5	6	7	8	9	10	11	12
fleet size	29	29	29	29	30	31	31	31	32	32	33	33
OP deviation	51	29	29	26	20	10	10	10	4	4	1	0
arcs#	222	244	244	248	255	263	263	263	270	270	277	275

Fig. 3 Varying weights in the objective function (4)

the historic schedule and more than half of the trains in *OP* can still remain over-provided.

(ii) Multi-type case

Experiments on problem instances with two unit types were also conducted. The same 156 trains in the North Area that are served by 33 Class 334 units are used for these tests. According to the operational rules, 24 of the 156 trains can be served by both Class 334 and Class 318, which cannot be coupled with each other. The other 132 trains can only be served by Class 334. Class 334 has a capacity of 183 seats while Class 318 has a capacity of 219 seats. The fleet size of Class 334 is limited to 30 units and for Class 318 is 3 units. A stopping criteria of 1 % relative gap in the BB tree and a maximum number of 5000 total BB tree nodes were set.

Similar patterns as for the single-type scenario can be observed in Table 3, such as that the fleet size tends to increase in order to reduce the *OP* deviation. The computational times were significantly larger for the multi-type scenario compared with the single-type scenario. Figure 4 graphically depicts the major results as shown in Table 3, in which it can be observed that the tendencies of the single-type case remain and are even more remarkable for the multi-type case.

5.2.2 Fixed fleet size

In order to obtain direct results on the maximum number of *OP* trains that can be achieved with a certain fleet size, we also performed experiments in which the

Table 3 Varying weights in the objective function with multiple types

C_2	Fleet size (318/334)	Arcs#	OP deviation	ECS#	BBNode#	Time (s)
0.01	29 (1/28)	248	31	1	482	181,299
0.02	29 (2/27)	246	30	1	51	4576
0.05	29 (1/28)	246	28	1	558	134,729
0.1	31 (1/30)	254	23	1	278	67,516
0.13	32 (2/30)	255	22	1	150	16,858
0.14	31 (2/29)	259	15	1	238	29,789
0.15	32 (3/29)	268	6	1	48	4097
0.16	32 (3/29)	268	6	2	31	2206
0.17	33 (3/30)	275	0	2	42	3478
0.18	33 (3/30)	275	0	2	46	3621
0.9	33 (3/30)	278	0	2	3166	6208

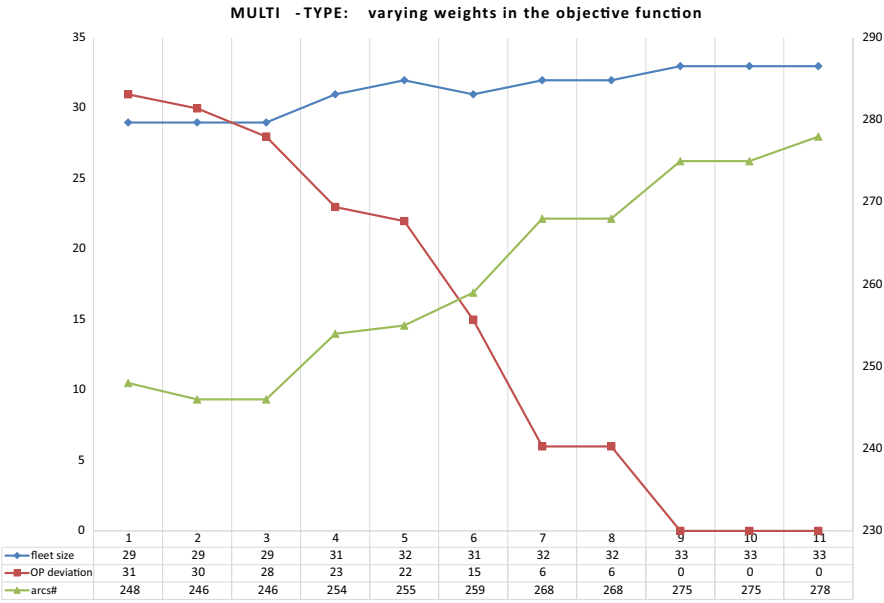


Fig. 4 Varying weights in the objective function (4) (multiple unit types)

deviation with respect to the *OP* trains is minimized while establishing an upper bound on the fleet size. From the historic schedules, it is known that one can over-provide the complete *OP* set with 33 train units. From the previous results, it is known that 29 train units are sufficient to meet all the target (lower level) capacity requirements. We conducted experiments within these fleet size bounds.

Results are presented in Table 4. As expected, the used fleet size equals the upper bound. For each value of the fleet size fixed, we obtain the best possible *OP* from the previous experiment.

Observe that the computation time tends to increase as the fleet size decreases. The reason is that the smaller the number of train units, the higher the difficulty of over-providing train capacity. For most of the fleet size values (from 31–33), the stopping criterion was that the gap is less than one OP train, thus yielding a strict optimal solution. However, when the fleet size is equal to 29 or 30, no optimal solution could be obtained by this stopping criterion. We have created other stopping criteria for these cases, by setting a maximum number of BB nodes of 2000 for the fleet size of 29, and 12000 for the fleet size of 30. In both cases, the resulting BB gap (the difference between the incumbent integer solution's objective value and the best BB tree lower bound) is equal to 2 OP trains.

5.3 Experiments on South area c314 instance

This group of experiments was conducted on the trains served by type c314 in the South area routes (see Fig. 2; Table 1). In the manual schedule, there are 9 OP trains and 12 UP trains compared with the PAX survey. Those UP trains mainly occur during peak hours. Due to the limited fleet size, some of these unsatisfied demand cannot be avoided. This is also apparent when the solver tries to solve the instance with the target (hard) capacity requirement as the PAX, it fails to give any feasible solution with the given fleet size upper bounds.

Let *UP* be the set of all UP trains in the c314 instance. Despite the fact that UP trains are unavoidable due to the limited unit resources ($|UP| \geq 1$), improvements can still be achieved. A proper subset of UP trains $SP \subset UP$ can be found where its member trains can have their capacity strengthened without violating the fleet size bound. Ideally, the operator would like to have $|SP|$ as large as possible since under-provision is a critical drawback for the solution quality. By using the bi-level capacity model (*P*), an optimal set *SP* can be determined.

This gives the principle of the experiments to be reported here. For the OP trains, $r_j = \rho_j^P < r'_j = \rho_j^H, \forall j \in OP$ was set. For the UP trains, $r'_j = \rho_j^P > r_j = \rho_j^H, \forall j \in UP$ was set. For the rest of the trains, where ρ_j^H is equivalent to ρ_j^P in terms of the number of units, only a target capacity requirement was set. Considering that even the default fleet size of 14 cannot meet all the PAX requirement, the fleet size is then fixed at 14, since either increasing or decreasing it would be meaningless. As

Table 4 Fixing scheduled fleet size from 29 to 33 train units

Fleet size	*OP*#	*OP* dev.	BB gap	Arcs #	ECS#	Time (s)
29	38	26	2	248	1	5916
30	46	18	2	255	1	146,997
31	54	10	0	263	1	40
32	60	4	0	270	1	37
33	64	0	0	276	2	35

Resulting number of elements in *OP*

minimizing the number of UP trains is the most critical task here and the fleet size as the most important part of the operational cost is already fixed, we solely minimized the deviation from the desirable level over the *OP* and *UP* trains in the objective, i.e., it was set as $\sum_{j\in OP\cup UP}|y_j - r'_j|$ and was labeled as "solver1". In addition, as here decreasing the number of UP trains in *UP* is more important than increasing the number of OP trains in *OP*, a second experiment was also carried out by only minimizing the deviation over *UP*, i.e. the objective was set as $\sum_{j\in UP}|y_j - r'_j|$ (labeled as "solver2").

The results on the two groups of experiments on the South area c314 routes are given in Table 5. The third column gives the number of OP trains within set *OP* while the fourth column gives the total number of OP trains. The former only shows the over-provided trains in the 64 OP trains (denoted as *OP*) in the original manual schedule; the latter gives the over-provided trains in all the 156 train services. However, UP trains can only appear within the set *UP* since the target level capacity requirements prevents the appearance of new UP trains. Both the two experiments were solved to optimality by branch-and-price. Compared with the operated schedule, the solution of solver1 maintains all the 9 OP trains in *OP* and adds four extra OP trains. In addition, it reduces the number of UP trains from 12 to 10, making a 16.7 % improvement without increasing the fleet size. As for solver2, since only the UP trains were considered, it can only maintain 3 OP trains out of the 9 in *OP* while it also adds 8 extra OP trains. So, its performance w.r.t. OP trains is worse than solver1's. On the other hand, solver2 succeeds in reducing the number of UP trains from 12 to 8, giving a 33 % improvement without increasing the fleet size.

The above results illustrate the advantage of the bi-level capacity model in better satisfying passenger demands, as well as keeping the pattern given by historic schedules.

6 Conclusions

We have introduced the train unit scheduling problem with bi-level, target and desirable, capacity requirements. For cases in which all the demand is satisfied, the first level concerns strict passenger capacity requirements, which should be strictly satisfied, and the second level concerns historic capacity provisions that will be satisfied as much as possible. For cases in which there exists unsatisfied demand, there exists some trains for which the target level corresponds to the historic

Table 5 Experiments on the south area c314 instance, fixed fleet size of 14

	Objective	OP# (*OP*)	OP# (all)	UP#	Time (s)
Manual	–	9	–	12	–
Solver1	$\sum_{j\in OP\cup UP}\|y_j - r'_j\|$	9	13	10	10,862
Solver2	$\sum_{j\in UP}\|y_j - r'_j\|$	3	11	8	6070

provision and others for which this corresponds to the PAX. In the railway context it is often required to maintain the historic pattern of unit resource distribution for OP trains wherever possible since this often contains implicit knowledge on agreements or expectations of transport authorities. Moreover, this helps to reinforce the robustness of the schedule with respect to changes in passenger demands.

We propose different strategies to deal with these two levels within the train unit scheduling optimization. Our methodology has been applied to real-world data provided by First ScotRail. It is shown that applying these strategies yields a set of efficient solutions, which in every case improves the existing schedule. With the proposed method, all the demand is satisfied with a 12 % smaller fleet size and nearly the 60 % most loaded train services within the over-provided ones in which the historic capacity provisions are maintained. In cases in which there exists unsatisfied demand, that is, under-provided trains, the proposed method reduces the under-provided trains by more than 33 % maintaining the same fleet size, thus reducing significantly the unsatisfied demand.

A byproduct considering different levels of capacity requirements is that future expected demand growth may also be considered. This is especially relevant in the context of franchise bidding, where future growth in passenger demands should be taken into consideration. In this context, multi-level capacity requirements would be useful for scheduling considerations. Further work is to develop a multi-level capacity requirements model taking all the relevant aspects of franchise bidding into account. In doing so, multicriteria optimization may also be considered at the desirable level.

Acknowledgments This research is supported by an EPSRC project EP/M007243/1. We would like to also thank First ScotRail for their kind and helpful collaboration and for providing us relevant data, part of which is commercially sensitive. The data that can be made publicly available is deposited in http://doi.org/10.5518/5.

References

Alfieri A, Groot R, Kroon LG, Schrijver A (2006) Efficient circulation of railway rolling stock. Transp Sci 40(3):378–391

Cacchiani V, Caprara A, Toth P (2010) Solving a real-world train-unit assignment problem. Math Program B 124(1–2):207–231

Cacchiani V, Caprara A, Toth P (2012) Models and algorithms for the train unit assignment problem. Combinatorial Optimization, vol 7422., Lecture Notes in Computer ScienceSpringer, Berlin, pp 24–35

Cacchiani V, Caprara A, Maróti G, Toth P (2013a) On integer polytopes with few nonzero vertices. Oper Res Lett 41(1):74–77

Cacchiani V, Caprara A, Toth P (2013b) A Lagrangian heuristic for a train-unit assignment problem. Discrete Appl Math 161(12):1707–1718

Cadarso L, Marín A (2011) Robust rolling stock in rapid transit networks. Comput Oper Res 38(8):1131–1142

Fioole PJ, Kroon L, Maróti G, Schrijver A (2006) A rolling stock circulation model for combining and splitting of passenger trains. Eur J Oper Res 174(2):1281–1297

Fuchsberger M, Lüthi PDHJ (2007) Solving the train scheduling problem in a main station area via a resource constrained space-time integer multi-commodity flow. Institute for Operations Research ETH Zurich

Jiang Z, Tan Y, Yalcinkaya O (2014) Scheduling additional train unit services on rail transit lines. Mathematical Problems in Engineering, vol 2014. http://dx.doi.org/10.1155/2014/954356

Kroon LG, Lentink RM, Schrijver A (2008) Shunting of passenger train units: an integrated approach. Transp Sci 42(4):436–449

Lin Z, Kwan RSK (2013) An integer fixed-charge multicommodity flow (FCMF) model for train unit scheduling. Electron Notes Discrete Math 41:165–172

Lin Z, Kwan RSK (2014) A two-phase approach for real-world train unit scheduling. Public Transp 6(1):35–65

Maróti G (2006) Operations research models for railway rolling stock planning, Ph.D. thesis. Eindhoven University of Technology, The Netherlands

Peeters M, Kroon LG (2008) Circulation of railway rolling stock: a branch-and-price approach. Comput OR 35(2):538–556

Schrijver A (1993) Minimum circulation of railway stock. CWI Q 6:205–217

Tracsis PLC (2015) TRACS-RS—rolling stock planning software.http://www.tracsis.com/software/tracs-rs (visited on 1 March 2015)